2К

УДК 004.934.1'1

***М.Х. Карабалаева*[1]*, А.В. Ниценко*[2]*, В.Ю. Шелепов*[3]**
[1]Евразийский национальный университет им. Л.Н. Гумилева, г. Астана, Казахстан
[2]Институт проблем искусственного интеллекта МОН Украины и НАН Украины, г. Донецк
[3]Государственный университет информатики и искусственного интеллекта,
г. Донецк, Украина
mkarabal@mail.ru, nav_box@mail.ru, shel@iai.dn.ua

Обнаружение и выделение звука [р] в речевом сигнале

Статья посвящена описанию двух новых алгоритмов обнаружения и выделения в речевом сигнале русского и казахского звука [р]. Оба алгоритма нацелены на выделение в сигнале низкоамплитудных участков, соответствующих моментам удара языка о небо. Один алгоритм оперирует численным аналогом полной вариации, другой – использует последовательные сглаживания и количество точек постоянства.

Целью работы является обнаружение в речевом сигнале фрагментов, соответствующих произнесенному звуку [р], и определение границ этих фрагментов. Идеология и методы настоящей работы лежат в сфере подходов к пофонемному распознаванию речи, подробное изложение которых содержится в книге [1].

1. Звук [**р**] – твердый, переднеязычный, сонорный, дрожащий согласный звук, очень распространенный как в русской, так и в казахской речи. В словах может встречаться в разных позициях, в сочетаниях с различными гласными и согласными звуками. При произнесении этого звука кончик языка вибрирует, под напором выдыхаемого воздуха, ударяя по небу, вследствие чего звук [**р**] получается дрожащим, раскатистым.

При проведении процедуры сегментации оцифрованного речевого сигнала, описанной в работе [2], звук [**р**], как правило, попадает в класс голосовых согласных. Однако особенности этого звука позволяют распознавать его в независимости от фонетического окружения, без необходимости предварительной сегментации.

Речевой сигнал, оцифрованный звукозаписывающим устройством, представляет собой массив отсчетов (сэмплов) x_i. Если взглянуть на речевой сигнал в амплитудно-временно́м представлении, сразу заметно, что на участках, соответствующих звуку [**р**], амплитуда сигнала резко падает там, где [**р**] ударяет по небу (рис. 1). На этих коротких участках падает и величина, называемая вариацией:

$$V = \sum_{i=1}^{n-1} \left| x_{i+1} - x_i \right| \quad (1)$$

– численный аналог полной вариации функции для дискретного случая.

Используем этот факт для обнаружения [**р**] в сигнале.

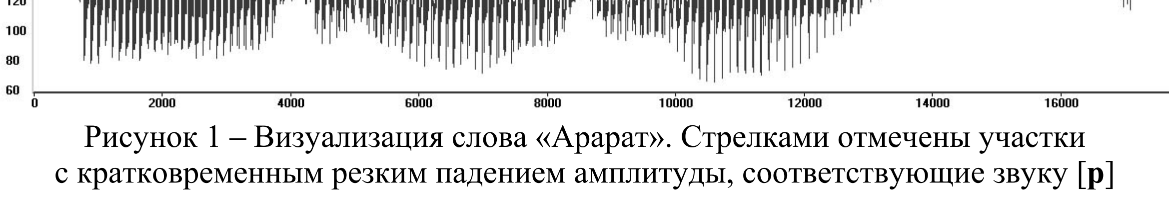


Рисунок 1 – Визуализация слова «Арарат». Стрелками отмечены участки с кратковременным резким падением амплитуды, соответствующие звуку [**р**]

Разобьем сигнал на последовательные окна по 128 отсчетов. В каждом окне вычислим вариацию (1), запишем полученные числа в массив. Затем поищем в этом массиве такие элементы V_k, что:

1) числовое значение элемента V_k попадает в заданный интервал: $a < V_k < b$;

2) максимальный из трех предшествующих элементов превышает данный элемент более, чем в c раз: $max(V_{k-1}, V_{k-2}, V_{k-3}) > c^* V_k$;

3) максимальный из трех последующих элементов также превышает данный элемент более, чем в c раз: $max(V_{k+1}, V_{k+2}, V_{k+3}) > c^* V_k$;

4) сумма максимального из трех предшествующих элементов и максимального из трех последующих элементов превышает данный элемент более, чем в d раз (причем $d > 2^*c$): $max(V_{k-1}, V_{k-2}, V_{k-3}) + max(V_{k+1}, V_{k+2}, V_{k+3}) > d^* V_k$; $d > 2^*c$.

Если в массиве найдутся элементы, отвечающие условиям (1 – 4), то будем считать, что они соответствуют участкам с вибрирующим [**р**].

Обратим внимание на тот факт, что при произнесении звука [**р**] кончик языка может ударить по небу не один раз, а несколько (например, казахское слово «бар» мы можем произнести раскатисто – «бар-р-р»). Тогда на графике речевого сигнала будет зафиксировано несколько кратковременных падений амплитуды, следующих друг за другом (рис. 2). Однако все они соответствуют одному и тому же звуку [**р**].

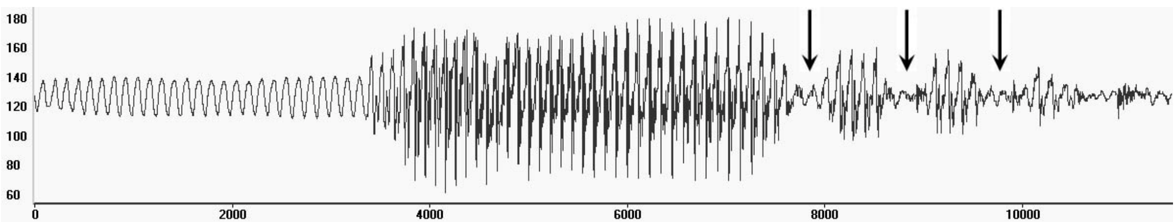


Рисунок 2 – Визуализация слова «бар-р-р». Стрелками отмечены участки с кратковременным резким падением амплитуды, соответствующие звуку [**р**]

В связи с этим введем ограничение длины для расстояния между соседними [**р**]:

5) если в массиве присутствуют 2 элемента V_p и V_q, отвечающих описанным условиям, то будем считать, что они соответствуют двум разным звукам [**р**], только в случае, когда они отстоят друг от друга более, чем на n позиций: $p - q > n$. В противном случае они соответствуют одному и тому же звуку [**р**].

Участок сигнала, содержащий элементы массива, описываемые условиями (1 – 5), пометим на графике сигнала меткой «R».

Тестирование описанного алгоритма выявило, что иногда на стыке голосового согласного и гласного звуков (например, в слове «бал» на стыке [**б**] и [**а**]) вариация также может кратковременно упасть относительно соседних участков, что приводит к возникновению лишней метки «R». Однако для голосовых согласных вариация в целом меньше, чем для [**р**]. Поэтому, чтобы избежать лишних меток в подобных случаях, добавим еще один (не относительный, а абсолютный) порог для «соседей» элемента V_k:

6) максимальный из трех предшествующих элементов превышает порог e: $max(V_{k-1}, V_{k-2}, V_{k-3}) > e$;

7) максимальный из трех последующих элементов также превышает порог e: $max(V_{k+1}, V_{k+2}, V_{k+3}) > e$.

Совокупность условий (1 – 7) позволяет нашей системе уверенно обнаруживать твердый звук [**р**] в разных позициях: в начале, в середине и в конце слова, в сочетаниях с гласными и согласными (рис. 3а, 3б, 3в).

Значения порогов для нашей системы: $a = 8$, $b = 50$, $c = 2.5$, $d = 6.5$, $e = 70$, $n = 3$.

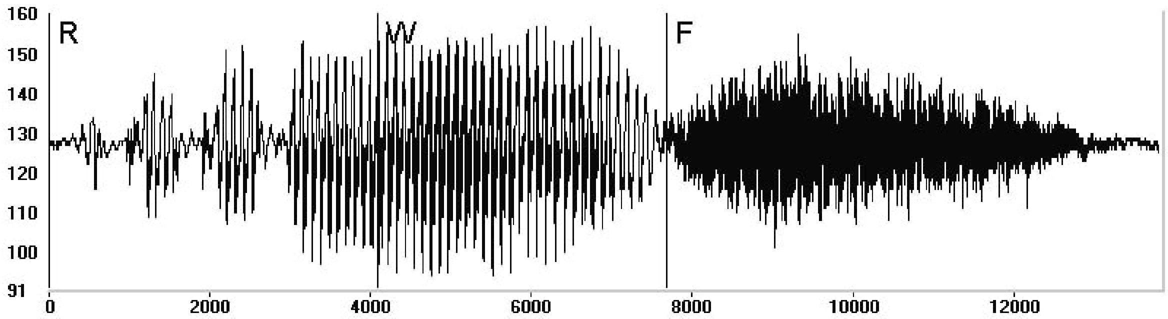


Рисунок 3а) – Сегментация казахского слова «рас».
Метки: R – звук [**р**], W – гласный звук, F – глухой фрикативный согласный звук

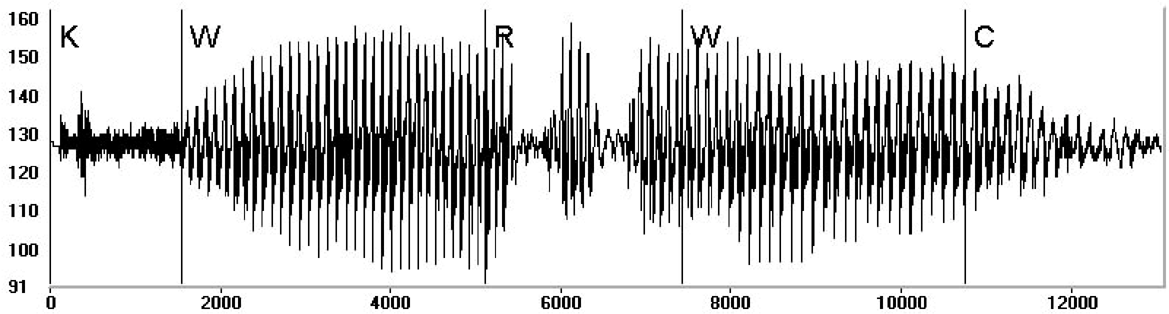


Рисунок 3б) – Сегментация казахского слова «қара». Метки: K – казахский звук [**қ**], W – гласный звук, R – звук [**р**], C – голосовая вставка в конце слова

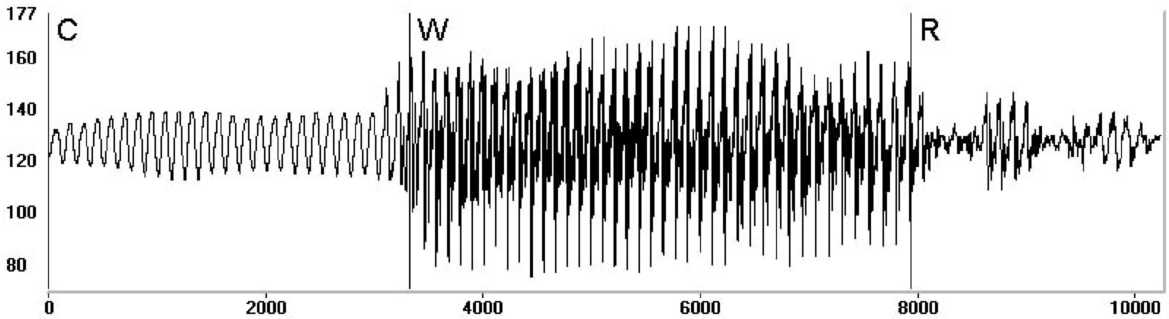


Рисунок 3в) – Сегментация казахского слова «бар».
Метки: C – голосовой согласный звук, W – гласный звук, R – звук [**р**]

2. Пусть имеется одномерный числовой массив и задан некоторый порог p. Построим символьную последовательность S, поставив в соответствие членам массива, которые больше p, символ «В» (выше порога), остальным символ «Н» (ниже порога). Будем называть эту процедуру, применяемую к числовому массиву, первичной «В-Н»-обработкой с порогом p.

Назовем сглаживанием сигнала

$$y_1, y_2, \ldots$$

обработку его 3-точечным скользящим фильтром

$$y_i = \frac{y_{i-1} + y_i + y_{i+1}}{3}.$$

Изложим еще один алгоритм детектирования и выделения звука «р». Он использует сглаживание и число точек постоянства, то есть таких моментов времени, что в следующий момент значение сигнала не меняется.

На рис. 4 приведено амплитудно-временное представление сигнала, соответствующего слову «сорока».

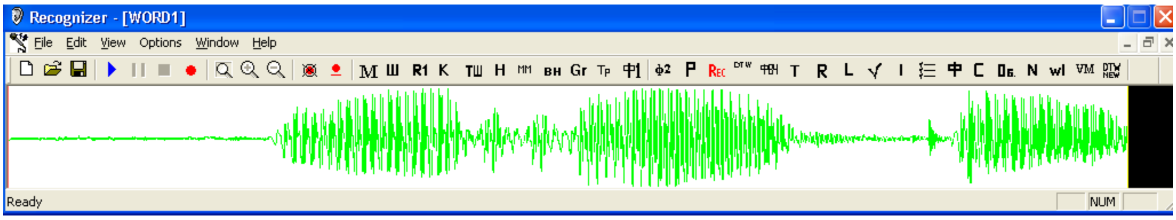


Рисунок 4 – Визуализация амплитудно-временного представления слова «сорока»

Рисунки 5 – 7 представляют результаты 10-кратного, 70-кратного и 100-кратного сглаживания исходного сигнала.

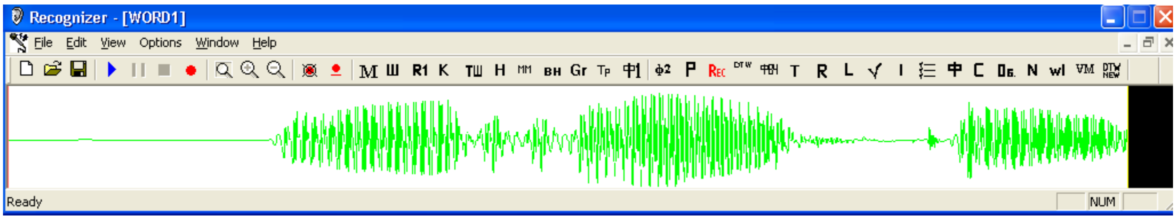


Рисунок 5 – Визуализация амплитудно-временного представления слова «сорока» после 10-кратного сглаживания

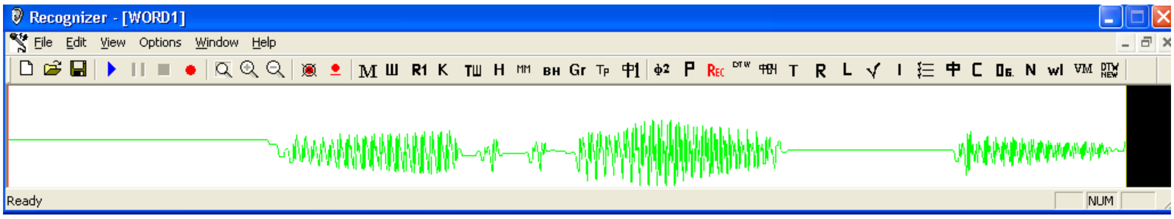


Рисунок 6 – Визуализация амплитудно-временного представления слова «сорока» после 70-кратного сглаживания

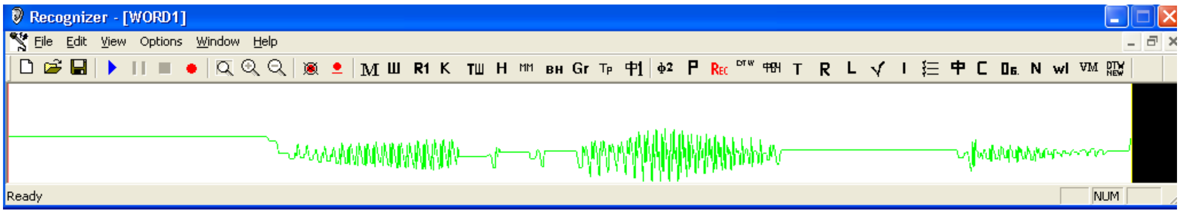


Рисунок 7 – Визуализация амплитудно-временного представления слова «сорока» после 100-кратного сглаживания

Из этих рисунков видно, что в случае достаточно большого числа сглаживаний на участках, соответствующих ударам языка по небу («р»-удар), число точек постоянства больше числа точек непостоянства. При этом длины этих участков относительно малы. Это позволяет, сделав первичную В-Н-обработку с порогом 0, выделять «р»-удары по этим признакам как достаточно короткие Н-участки. Однако при этом в число выделенных могут попасть участки других голосовых (и глухих) звуков. Но другие звуки, в отличие от [р], являются достаточно однородными. Поэтому образовавшиеся там Н-участки быстро расширяются с увеличением числа сглаживаний. В то же время выделенные участки «р»-ударов остаются короткими при некотором увеличении числа сглаживаний. Учет этого позволит избавиться от выделенных участков, не относящихся к «р». Итак, вопрос должен решаться, если мы сумеем формализовать «историю» того, что получается при некотором числе последовательных сглаживаний сигнала.

Суммируя сказанное, мы приходим к алгоритму, который реализуется с помощью двумерного массива, иллюстрируемого таблицами на рис. 9 и 10 .

Пусть для примера анализируется сигнал рис. 8, отвечающий слову «пара».

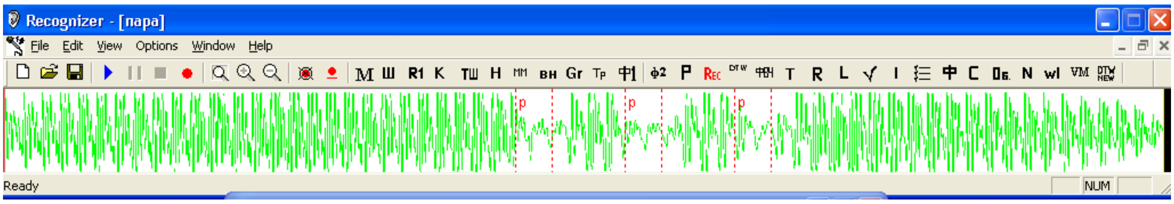


Рисунок 8 – Визуализация амплитудно-временного представления слова «пара»; «р»-удары выделены по описываемому алгоритму

Сигнал разбивается на неперекрывающиеся окна по 256 отсчетов. На каждом из них вычисляется разность между количеством точек непостоянства и количеством точек постоянства. Полученный числовой массив подвергается первичной «В-Н»-обработке с порогом 0 (при необходимости этот порог может быть заменен другим).

Столбцы таблицы на рис. 9 отражают результаты такой обработки сигнала после определенного числа сглаживаний. Первый вертикальный столбец – после 10-кратного сглаживания сигнала. Второй – после 20-кратного сглаживания и так далее. Всего здесь выполнено 15 последовательных 10-кратных сглаживаний.

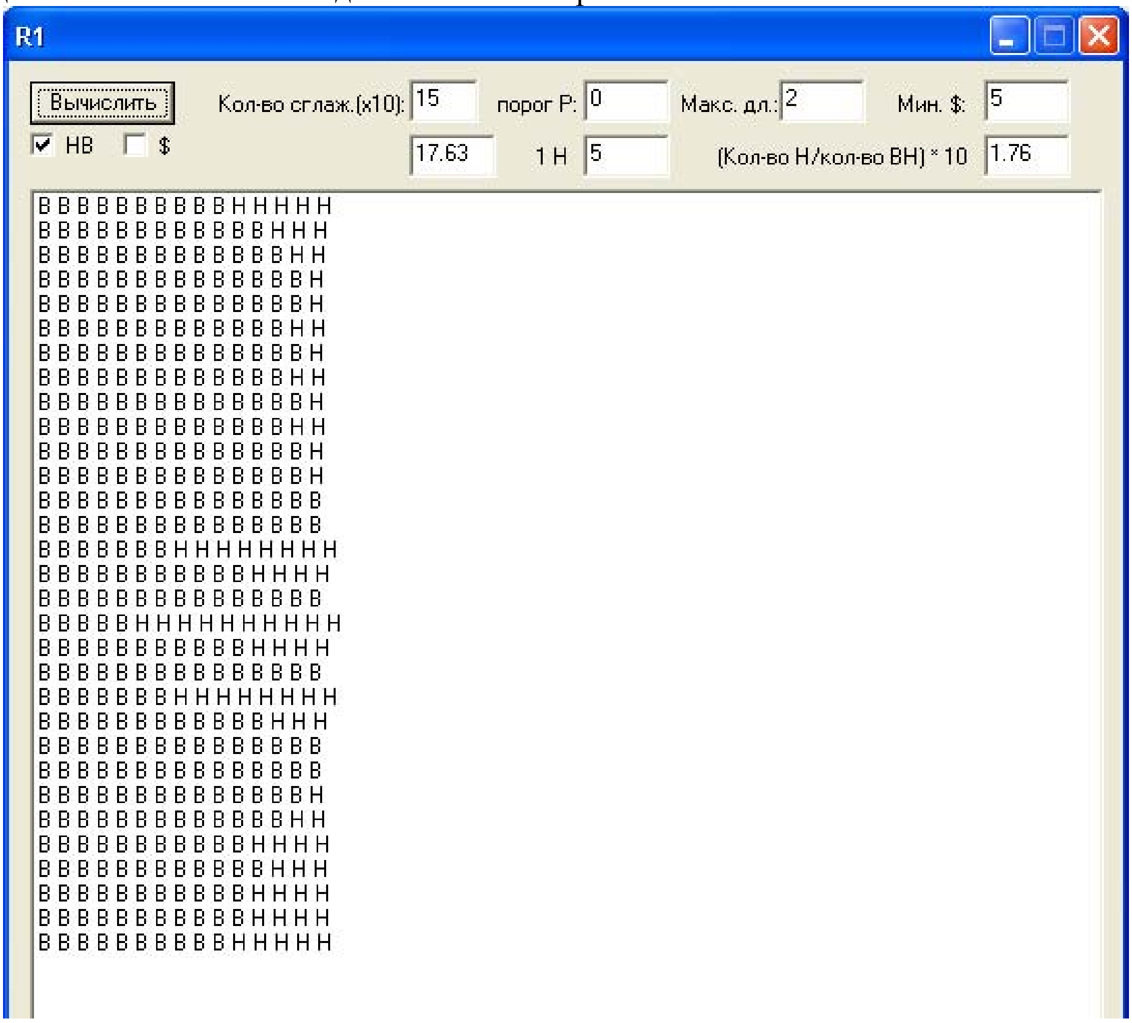


Рисунок 9 – Двумерный массив символов, полученный вышеописанным образом

Таблица на рис. 10 возникает в результате выделения в столбцах предыдущей таблицы последовательностей «Н», длины которых не превосходят 2. В них символ «Н» заменяется значком «$» (в нашем примере более длинных Н-последовательностей не оказалось). Наконец, анализируя строки полученной таблицы, выделяем те из них, где количество рядом стоящих «$» не меньше 5. Этим строкам соответствуют участки «р»-ударов.

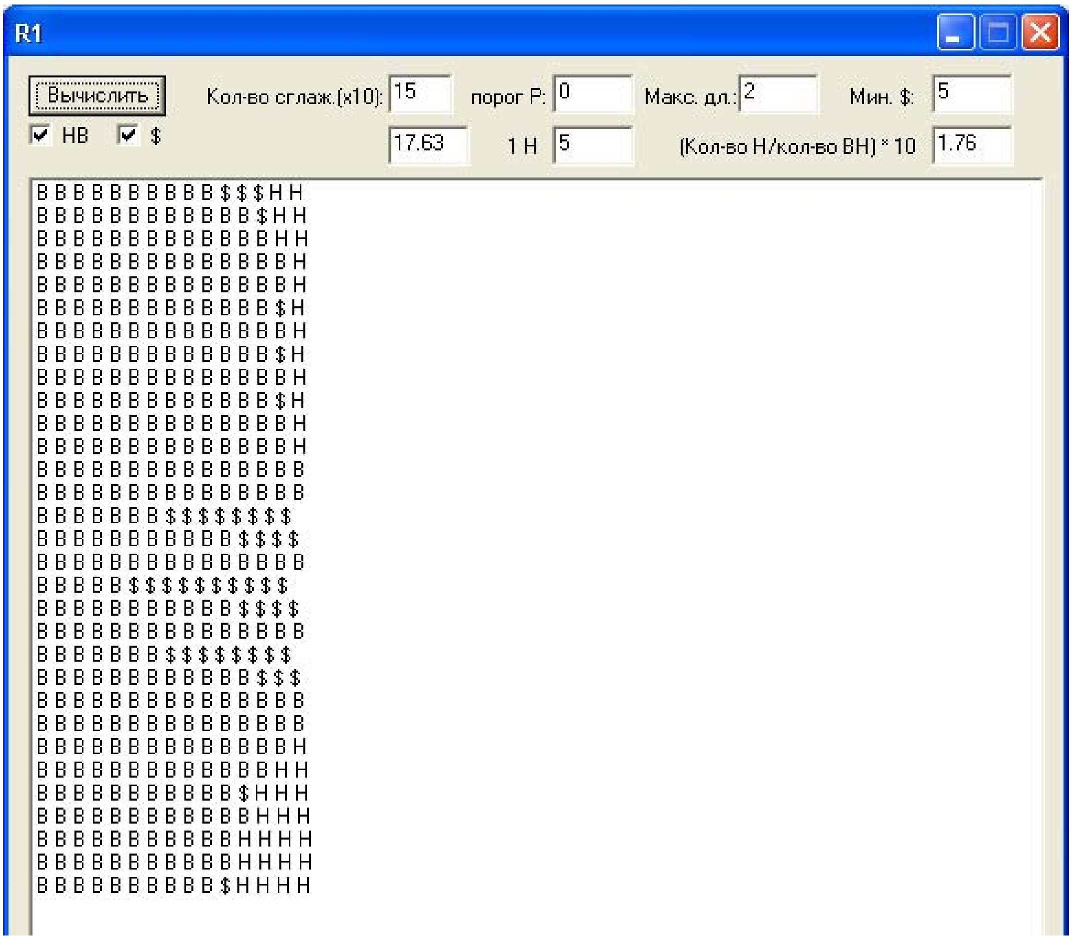


Рисунок 10 – Окончательная таблица, используемая для детектирования «р»-ударов

Мы ограничиваемся таким наглядным описанием алгоритма и позволим себе не приводить его формального описания. В общем случае алгоритм содержит 4 параметра: число 10-кратных сглаживаний, порог первичной «В-Н»-обработки, максимальная допустимая длина Н-отрезка в столбце, минимальное количество идущих подряд символов $ в строке.

Так же, как и ранее изложенный, только что описанный алгоритм, успешно находит твердое «р» в позициях начала, середины и конца слова в любом фонетическом окружении. Он справляется также с мягким [р].

Пример:

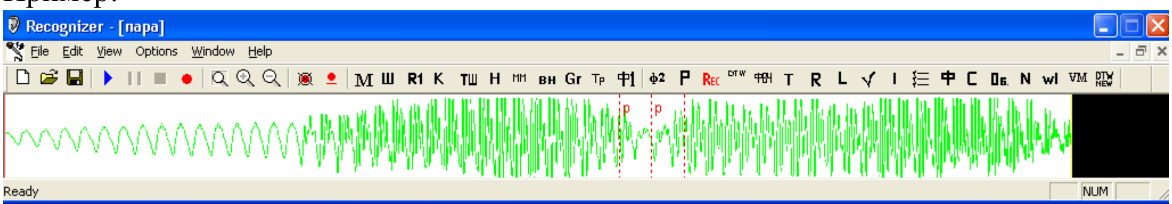


Рисунок 11 – Визуализация амплитудно-временного представления слова «море»

Начало и конец участка «р»-удара отмечаются вертикальными метками. Изложенное относится к проблеме обнаружения (детектирования) «р» в речевом сигнале. Наши подходы к пофонемному распознаванию основаны на предварительной сегментации – разбиении речевого сигнала на участки гласных (обозначение сегмента W), голосовых согласных (обозначение сегмента C), глухих фрикативных звуков (обозначение сегмента F) и аффрикат, глухих взрывных (паузообразных) звуков (обозначение сегмента P). Об алгоритмах такой сегментации (будем называть ее основной) см. работу [2]. Теперь мы хотим добавить в число выделяемых отрезков отрезки «р».

После завершения основной сегментации в нее добавляется информация об «р» по следующим правилам:

– если хотя бы одна из полученных ранее меток для «р» попадает в сегмент «C», то весь этот сегмент помечается как «р»;

– если метка для [р] попадает в сегмент «W», то участок от этой метки до начала следующей фонемы помечается как «р»;

– если метки для [р] попадают в соседние сегменты «W» и «C», то участок от первой метки [р] в сегменте «W» до конца сегмента «C» помечается как «р».

Литература

1. Шелепов В.Ю. Лекции о распознавании речи / Шелепов В.Ю. – Донецьк : ІПШІ «Наука і освіта». – 192 с.
2. Шелепов В.Ю. Структурная классификация слов русского языка. Новые алгоритмы сегментации речевого сигнала и распознавания некоторых классов фонем / В.Ю. Шелепов, А.В. Ниценко // Искусственный интеллект. – 2007. – № 1. – С. 213-224.

М.Х. Карабалаєва, А.В. Ніценко, В.Ю. Шелєпов

Виявлення і виділення звуку [р] у мовному сигналі

Стаття присвячена опису двох нових алгоритмів виявлення і виділення у мовному сигналі російського та казахського звуку [р]. Обидва алгоритми націлені на виділення в сигналі низькоамплітудних ділянок, які відповідають моментам удару язика об піднебіння. Один алгоритм оперує чисельним аналогом повної варіації, інший – використовує послідовні згладжування і кількість точок сталості.

M.H. Karabalajeva, A.V. Nicenko, V.Ju. Shelepov

Detection and Isolation of the Phoneme [r] in the Speech Signal

This paper describes two new algorithms for detection and isolation of the phoneme [r] in the Russian and Kazakh speech signal. Both algorithms aim to detect a low-amplitude signal frames corresponding to the moments of interaction between the tongue and the hard palate. One algorithm operates with a numerical analog of the full variation, the other uses sequential smoothing and the number of points of constancy.

Статья поступила в редакцию 03.03.2011.